THIS PAGE BLANK (USPTO)

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : C12Q 1/68	A1	(11) International Publication Number: WO 00/22173
		(43) International Publication Date: 20 April 2000 (20.04.00)

(21) International Application Number: PCT/US99/24388

(22) International Filing Date: 14 October 1999 (14.10.99)

(30) Priority Data:
09/174,364   15 October 1998 (15.10.98)   US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US   09/174,364 (CON)
Filed on   15 October 1998 (15.10.98)

(71) Applicant *(for all designated States except US)*: PRINCETON UNIVERSITY [US/US]; Office of Technology & Trademark Licensing, New South Building, 5th floor, P.O. Box 36, Princeton, NJ 08544–0036 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: LEVINE, Arnold, J. [US/US]; 138 Fitzrandolph Road, Princeton, NJ 08540 (US). ALON, Uri [IL/US]; Hibben 7L, Princeton, NJ 08540 (US).

(74) Agents: DURDIK, Paul, A. et al.; Townsend and Townsend and Crew LLP, 8th Floor, Two Embarcadero Center, San Francisco, CA 94111–3834 (US).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: QUANTITATIVE ANALYSIS OF HYBRIDIZATION PATTERNS AND INTENSITIES IN OLIGONUCLEOTIDE ARRAYS

(57) Abstract

Systems and methods for enhanced quantitative analysis of hybridization intensity measurements obtained from oligonucleotide probes and other probes exposed to target samples are provided by virtue of the present invention. One embodiment ameliorates the effects of high frequency noise superimposed on a hybridization intensity measurement signal measured over successive probe alignments to a target sample sequence. Detection of expressed genes and ESTs and quantitative measurement of expression level may be improved. Mutation detection and base calling may be improved.

QUANTITATIVE ANALYSIS OF HYBRIDIZATION PATTERNS AND INTENSITIES IN OLIGONUCLEOTIDE ARRAYS

BACKGROUND OF THE INVENTION

The present invention relates to computer systems and more particularly to systems and methods for analysis of hybridization of samples to oligonucleotide probes or other polymer probes.

Devices and computer systems for forming and using arrays of materials on a substrate are known. The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. See U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid). It is also possible to employ other types of probes or probes that are not included in arrays or chips.

Such probes are used for, e.g, base calling, detection of mutations, and analysis of gene expression. For all of these objectives, a typical technique is to expose the probes to target nucleic acid samples that have been marked with fluorescent or otherwise radioactive labels. For each probe or group of probes, a hybridization intensity is determined based on observed fluorescence or radioactivity. The hybridization intensity may also be measured in some other way.

These hybridization intensities are the basis for further analysis including base calling, mutation detection, and evaluation of expression of genes or expressed sequence tags. See European Patent Office Publication No. 0717113A and European Patent Office Publication No. 0848067, the contents of both publications being incorporated herein by reference.

Expression evaluation makes use of hybridization intensities determined from pairs of probes where each pair includes a perfect match probe and a mismatch

2

probe. The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular subsequence of a sequence of interest in a target nucleic acid. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a particular target

5   sequence.

For example, to determine the concentration of a particular mRNA sequence indicative of expression of a gene or EST of interest, a series of pairs of perfect match and mismatch probes may be provided. Each pair may include a perfect match probe perfectly complementary to a subsequence of interest. The mismatch probe may

10   differ in one position from the perfect match probe. Each probe may include a series of e.g., 25 bases. The mRNA sequence may be interrogated by a series of probe pairs having successive alignments to the mRNA sequence.

After hybridization intensities are obtained, the number of instances of when the perfect match intensity is greater than the mismatch intensity is obtained, along

15   with the average of the logarithm of the perfect match to mismatch ratios for all the probe pairs. To determine the quantitative abundance of mRNA, the average of the difference between perfect match and mismatch hybridization intensity is also computed.

Further opportunities exist, however, to improve the accuracy of assessments of expression levels. High frequency noise can result from variations in

20   probe alignment to mRNA sequences, causing hybridization intensity to exhibit spurious peaks rather than smooth variation. This high frequency noise is especially prevalent in array designs where there are relatively small number of probes per gene and therefore less opportunity to average out the high frequency noise over results from large number of probes.

25   What is needed are systems and methods for reducing the deleterious affects of the high frequency noise found in the hybridization intensity measurements.

SUMMARY OF THE INVENTION

Systems and methods for enhanced quantitative analysis of hybridization

30   intensity measurements obtained from oligonucleotide probes and other probes exposed to target samples are provided by virtue of the present invention. One embodiment ameliorates the effects of high frequency noise superimposed on a hybridization intensity signal measured over successive probe alignments to a target sample sequence. Detection

of expressed genes and ESTs and quantitative measurement of expression level may be improved. Mutation detection and base calling may be improved.

A nonlinear lowpass filter may be used to remove the effects of spurious peaks in this signal. Also, a hybridization spectrum including the hybridization intensities

5   measured over a series of probes may be compared to a reference hybridization spectrum to obtain a measure of similarity. The measure of similarity may indicate expression or non-expression of a particular gene or EST, or a point mutation.

In accordance with a first aspect of the present invention, a method for analyzing a nucleic acid sequence includes: inputting a plurality of hybridization

10   intensities of probes exposed to the sample nucleic acid sequence, and applying a non-linear filter to the plurality of hybridization intensities.

In accordance with a second aspect of the present invention, a method for analyzing a sample nucleic acid sequence includes: inputting a plurality of hybridization intensities of probes exposed to the sample nucleic acid sequence, the plurality of

15   hybridization intensities forming a hybridization spectrum of the sample nucleic acid sequence, and comparing the hybridization spectrum of the sample nucleic acid sequence to a reference hybridization spectrum to obtain an indication of similarity.

A further understanding of the nature and advantages of the inventions herein may be realized by reference to the remaining portions of the specification and the

20   attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention.

25   Fig. 2 shows a system block diagram of a typical computer system.

Fig. 3 is a flowchart describing steps of analyzing hybridization data using a non-linear filter according to one embodiment of the present invention.

Figs. 4A–4B depict the effects of low-pass filtering of hybridization data according to one embodiment of the present invention.

30   Fig. 5 is a flowchart describing steps of determining expression by use of hybridization spectra according to one embodiment of the present invention.

4

DESCRIPTION OF THE SPECIFIC EMBODIMENTS

Fig. 1 illustrates an example of a computer system that may be used to execute software embodiments of the present invention. Fig. 1 shows a computer system 1 which includes a monitor 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons such as mouse buttons 13. Cabinet 7 houses a CD-ROM drive 15 and a hard drive (not shown) that may be utilized to store and retrieve software programs including computer code incorporating the present invention. Although a CD-ROM 17 is shown as the computer readable medium, other computer readable media including floppy disks, DRAM, hard drives, flash memory, tape, and the like may be utilized. Cabinet 7 also houses familiar computer components (not shown) such as a processor, memory, and the like.

Fig. 2 shows a system block diagram of computer system 1 used to execute software embodiments of the present invention. As in Fig. 1, computer system 1 includes monitor 3 and keyboard 9. Computer system 1 further includes subsystems such as a central processor 50, system memory 52, I/O controller 54, display adapter 56, removable disk 58, fixed disk 60, network interface 62, and speaker 64. Removable disk 58 is representative of removable computer readable media like floppies, tape, CD-ROM, removable hard drive, flash memory, and the like. Fixed disk 60 is representative of an internal hard drive or the like. Code to implement aspects of the present invention may be operably disposed in or stored on any type of storage medium.

Other computer systems suitable for use with the present invention may include additional or fewer subsystems. For example, another computer system could include more than one processor 50 (i.e., a multi-processor system) or memory cache.

Arrows such as 66 represent the system bus architecture of computer system 1. However, these arrows are illustrative of any interconnection scheme serving to link the subsystems. For example, display adapter 56 may be connected to central processor 50 through a local bus or the system may include a memory cache. Computer system 1 shown in Fig. 2 is but an example of a computer system suitable for use with the present invention. Other configurations of subsystems suitable for use with the present invention will be readily apparent to one of ordinary skill in the art. In one embodiment, the computer system is an IBM compatible personal computer.

The VLSIPS™ and GeneChip™ technologies provide methods of making and using very large arrays of polymers, such as nucleic acids, on very small chips. See

5

U.S. Patent No. 5,143,854 and PCT Patent Publication Nos. WO 90/15070 and 92/10092, each of which is hereby incorporated by reference for all purposes. Nucleic acid probes on the chip are used to detect complementary nucleic acid sequences in a sample nucleic acid of interest (the "target" nucleic acid).

5   It should be understood that the probes need not be nucleic acid probes but may also be other polymers such as peptides. Peptide probes may be used to detect the concentration of peptides, polypeptides, or polymers in a sample. The probes must be carefully selected to have bonding affinity to the compound whose concentration they are to be used to measure.

10   In one embodiment, the present invention provides methods of analyzing information relating to the concentration of compounds in a sample as measured by binding of the compounds to polymers such as polymer probes. In a particular application, the concentration information is generated by analysis of hybridization intensity files for a chip containing hybridized nucleic acid probes. The hybridization of

15   a nucleic acid sample to certain probes may represent the expression level of one more genes or expressed sequence tags (ESTs). The expression level of a gene or EST is herein understood to be the concentration within a sample of mRNA or protein that would result from the transcription of the gene or EST.

          Concentration of compounds other than nucleic acids may be analyzed

20   according to one embodiment of the present invention. For example, a probe array may include peptide probes which may be exposed to protein samples, polypeptide samples, or peptide samples which may or may not bond to the peptide probes. By appropriate selection of the peptide probes, one may detect the presence or absence of particular proteins, polypeptides, or peptides which would bond to the peptide probes.

25   A system that designs a chip mask, synthesizes the probes on the chip, labels nucleic acids from a target sample, and scans the hybridized probes is set forth in U.S. Patent No. 5,571,639 which is hereby incorporated by reference for all purposes.

          The term "perfect match probe" refers to a probe that has a sequence that is perfectly complementary to a particular target sequence. The test probe is typically

30   perfectly complementary to a portion (subsequence) of the target sequence. The term "mismatch control" or "mismatch probe" refer to probes whose sequence is deliberately selected not to be perfectly complementary to a subsequence of a particular target sequence. For each mismatch (MM) control in an array there typically exists a

corresponding perfect match (PM) probe that is perfectly complementary to the same subsequence of a particular target sequence.

One possible probe selection strategy is to choose the PM probes to be perfectly complementary to successive subsequences of the target mRNA sequence. For example, the target sequences may be hundreds or thousands of bases long. Each perfect match probe may be 20-45 bases long. For example, in one such scheme, each probe is a 25-mer probe, i.e., the probes are 25 bases long. There may be probe pairs corresponding to every alignment to the target sequence, or there may be, e.g., 2-5 base pair differences in alignment for successive probe pairs. Also, for each alignment used there may be multiple probe pairs.

Hybridization intensities may be obtained by fluorescent scanning. The expression evaluation techniques, described for example in European Patent Office Publication No. 0848067 are based on relative measurements of the hybridization intensities for PM and MM probes. For example, the determination of whether the gene or EST is in fact expressed in the sample may be based on the number of probe pairs where the PM intensity exceeds the MM intensity by a threshold along with the average logarithm of the PM/MM ratios. Other criteria may be the number of probe pairs where both the PM intensity exceeds the MM intensity by a difference threshold or where the PM intensity divided by the MM intensity exceeds a ratio threshold. The quantitative expression level may depend on the average difference between PM and MM intensities. Hybridization intensities are the basis for all of these techniques.

For probe having successive alignments to a target sequence, the hybridization intensity will not typically vary smoothly but will rather exhibit spurious peaks. The present invention provides systems and methods for alleviating the deleterious effects of the peaks. In one embodiment, a nonlinear filter is applied to the hybridization data to remove these peaks.

Fig. 3 is a flowchart describing steps of processing hybridization measurements using a non-linear filter according to one embodiment of the present invention. The procedure of Fig. 3 may be applied, e.g., to the PM hybridization intensities, to the MM hybridization intensities, to the differences between PM and MM intensities for successive probes, to the ratios of PM and MM intensities for successive probes, or any combination of these measurements.

At step 302, the procedure accepts as input intensity measurements from probes used to detect the presence of a particular hybridized sequence. These measurements may be the perfect match measurements, mismatch measurements, match vs. mismatch difference measurements, ratio measurements, etc. The procedure is then applied for each alignment of probe to target sequence. Each alignment may be referred to as a "site" referring, e.g., to the base on the target that is complementary to a center base of the PM and MM probes. For a currently processed site, at step 304, the procedure isolates the hybridization intensity measurements collected from probes aligning to the target within a window of N sites surrounding the current site. If the current site is less than N/2 away from the beginning or end of the target sequence, the vector of the intensity measurements of the target sequence may be 'padded' by adding interpolated values to its beginning and end. For example, in an embodiment using a linear interpolation, if N=5 and each base i has a corresponding intensity measurement denoted as $X(i)$ for i ranging from 1 to N, the padded values are $2*X(1)-X(3)$ and $2*X(1)-X(2)$ followed by the sequences, $X(1)$ to $X(N)$, followed by padded values $2*X(N)-X(N-1)$ and $2*X(N)-X(N-2)$. Note that if the probe selection scheme provides for successive probes that may vary in alignment by more than one base, the N sites will not always be contiguous ones. If more than one probe pair has been used for each alignment, averages, medians, etc. may substitute for measurements obtained from one probe or probe pair.

At step 306, the intensity measurements obtained from each of the sites within the window of N sites are ranked in order of intensity. At step 308, the measurements from the M center sites are preserved and the rest are discarded, thus eliminating outliers. Steps 306 and 308 implement one type of lowpass nonlinear filter that may be used according to the present invention. Those of skill in the art will, however, appreciate that many nonlinear filters may be employed beneficially.

At step 310, the remaining M intensity measurements are averaged together. The resulting average replaces the original intensity measurement for the site at step 312. Alternatively, at step 310A, the procedure obtains the median of the remaining M intensity measurements. Then at step 312A, the resulting median replaces the original intensity measurement of the site. Processing of the current site then being complete, the procedure continues to the next site at step 314. Steps 304 through 314 then repeat for each succeeding alignment of probe to target sequence.

8

Effectively, this filter averages over the local intensity, throwing out the points with the highest and lowest intensities. Using such a filter on both the PM and MM signals, peaks and rapidly oscillating noise may be substantially reduced. It has been found that in a chip having about 100 probes per gene, the number of expressed genes

5   that may be unambiguously detected (in which most of the probes have PM>MM) increases from about 10% using the unfiltered hybridization intensity measurements to about 20% using the filtered hybridization intensity data.

Figs. 4A-4B depict the effects of filtering. Fig. 4A is a plot of unfiltered hybridization intensity as measured for various alignments to the target sequence. The

10   perfect match and mismatch intensities are plotted separately. Fig. 4B shows a plot of hybridization intensity after filtering according to the present invention. Again, the perfect match and mismatch intensities have been plotted separately. Note that a spurious mismatch peak 402 in the unfiltered plot is removed in the filtered plot. Also, in the filtered data, 72% of the probe pairs exceed the PM - MM threshold and the total

15   difference between PM and MM intensity is 2300. By contrast, in the unfiltered data only 55% of the probe pairs exceed the PM-MM threshold and the total intensity difference is only 200. Here, filtering makes the difference between detecting and not detecting gene expression.

The present invention also provides an even more sensitive system and

20   method for detection of gene expression. This high-sensitivity detection technique takes advantage of the property that each gene has a unique recurring pattern of hybridization intensity as evaluated over probe alignment. The pattern holds over disparate tissue types, including, e.g., ovarian and breast tumors, pre- and post-nude mouse cloning, and normal tissues. These patterns, herein referred to as gene hybridization spectra, are

25   thought to be due to changes in the hybridization efficiency resulting from variations in probe sequence. This gene hybridization spectrum may be understood as a distinct signature of each gene.

Fig. 5 is a flowchart describing steps of determining an expression using hybridization spectra according to one embodiment of the present invention. At step 502,

30   probes are selected for optimal detection of gene expression. Details of step 502 will be discussed more fully below. At step 504, a hybridization spectrum may be formed from intensity measurements for probes from a particular experiment. In the presently preferable embodiment, intensities are first filtered in accordance with the steps depicted

in Fig. 3 in order to remove spikes, prior to forming the hybridization spectrum. In one

embodiment, this spectrum includes the intensity measurements for successive PM probes

complementary to successive mRNA subsequences of the sequence along the gene. MM

probe measurements may also be included or one may use difference or ratio

measurements for successive probe pairs. The presently preferable embodiment forms

the hybridization spectrum using the intensity difference between PM and MM. The

hybridization spectrum may also be obtained by averaging intensities over many sets of

identical probes on the same chip or by averaging intensities obtained from many chips.

At step 506, this hybridization spectrum is compared to a reference

hybridization spectrum to determine whether or not a given gene or EST has been

expressed. The reference hybridization spectrum may be the hybridization spectrum

formed from intensity measurements on probes that have been exposed to a sample that is

known to include mRNA indicative of gene expression. Alternatively, the reference

hybridization spectrum may represent an average of measurements made on many

samples known to have the expressed gene. The comparison may be to a library of

reference hybridization spectra for different genes so that one experiment may be used to

measure expression of many genes.

In one embodiment, a result of the comparison is a first numerical

indicator of similarity between the newly formed hybridization spectrum and the

reference spectrum. A second numerical indicator may give a measure of the ratio of the

level of abundance of the mRNA in the new experiment to the level in the reference.

Any pattern matching algorithm can be used to perform the comparison.

In one embodiment, linear regression is used. Assume that Y is the newly formed

hybridization spectrum and that X is a reference hybridization specturm. The linear

regression algorithm finds the best linear relation between the signals, $Y = a*X + b$. Here

'a' is a linear fit coefficient that gives the ratio of the level of abundance of mRNA in the

new experiment to the level of the reference.

The linear regression algorithm further gives a regression coefficient r,

which has a value between -1 and 1. A magnitude of r being close to 1 means a perfect

linear correlation between the two spectra. When r is close to zero, that means the two

spectra are completely uncorrelated. The regression coefficient thus serves as an

indicator of whether a particular gene or EST is expressed.

10

It has been found that for a gene chip having about 100 probe pairs per gene, comparing different genes in the same experiment gives a regression coefficient of less than 0.2, indicating that the hybridization spectra of different genes are uncorrelated. When comparing the hybridization spectra of the same gene from different experiments, one obtains regression coefficients of greater than 0.8 for more than 90% of the genes. The increased sensitivity of this technique is due to the fact that the comparison takes into account all of the intensity information and not just a mean difference or some other value that attempts to represent the intensity results for all of the probes.

One may employ this hybridization spectrum evaluation technique to evaluate the quality of cDNA libraries by comparing the hybridization spectra of cDNA samples to reference spectra taken from samples of known quality. Another application would be to compare spectra to detect mutations or call bases. The reference hybridization spectrum would represent the wild type. Localized differences between the reference hybridization spectrum and the hybridization spectrum from a new experiment would represent point mutations. By comparing a new hybridization spectrum based on sample having, e.g., one unknown base, to four reference spectra collected from samples having each of the four possible bases at that position, one can call the base based on the closest matching of the four reference spectra. Again, the matching here may be based on a measure of localized differences such as e.g., mean square error, rather than an overall linear regression. To measure local differences one can perform the linear regression procedure over a small section of spectrum corresponding to the mutation point.

A modification of the above technique would be to group genes or ESTs together into families based on their hybridization spectra. Hybridization spectra are formed based on samples that express a known assortment of genes or ESTs. Those spectra that correlate with each other closely based on any pattern matching technique including the linear regression procedure outlined above are designated to be part of the same family. A family here is a group of genes or ESTs that have similar hybridization spectra.

At step 502, the design of probe arrays and probe selection strategies may be optimized to take advantage of the hybridization spectrum approach to detecting gene expression. A goal is to provide a sufficient number of probes so that each gene tested by a given array will have a detectable unique hybridization spectrum while maximizing the number of genes detectable with the available probes on an array.

11

Within the spectrum for a particular gene, each probe may be thought to have either a high (H) or low (L) hybridization efficiency. Thus a spectrum may be expressed as a string of H's and L's, e.g., "HLLHHHHL... ." If there are k probes per gene, there are 2^k distinct sequences of high and low intensities. Thus, for N genes, it

5   may be sufficient to have of order of k=log₂N probes per gene in order to have a detectably distinct pattern for each gene. Of course, the probes should then be selected to give a unique pattern for the gene.

Table 1 gives a minimum number of probes per gene for various numbers of genes in an array.

10

Number of Genes in Array	Number of Probes per Gene
10^2	7
10^3	10
10^4	14
10^5	17
10^6	20

Table 1

Adding more probes per gene will improve performance but only up to a point beyond which addition of further probes per gene does not add performance but

15   only reduces the number of genes that a given array can detect.

In the foregoing specification, the invention has been described with reference to specific exemplary embodiments thereof. It will, however, be evident that various modifications and changes may be made thereunto without departing from the broader spirit and scope of the invention as set forth in the appended claims and their full

20   scope of equivalents. For example, it will be understood that wherever "expression level" is referred to, one may substitute the measured concentration of any compound. Also, wherever "gene" is referred to, one may substitute the term "expressed sequence tag."

WHAT IS CLAIMED IS:

1   1.   In a computer system, a method for analyzing a sample nucleic
2   acid sequence, said method comprising:
3   inputting a plurality of hybridization intensities of pairs of perfect match
4   and mismatch probes, each pair including a perfect match probe perfectly complementary
5   to a particular nucleic acid subsequence indicative of expression of a gene or EST and a
6   mismatch probe having at least one base mismatch with said particular subsequence;
7   evaluating relative hybridization intensities for a plurality of subsequences
8   by comparison of hybridization intensities between perfect match probes perfectly
9   complementary to said subsequences and mismatch probes having at least one base
10   mismatch with said particular subsequences; and
11   applying a non-linear filter to said relative hybridization intensities

1   2.   The method of claim 1 wherein the applying comprises:
2   for a particular base position, averaging relative hybridization intensities
3   over subsequences aligned to said particular base position and subsequences aligned to
4   surrounding base positions, excluding subsequences having outlying relative
5   hybridization intensities.

1   3.   The method of claim 1 wherein said relative hybridization
2   intensities are evaluated as ratios of perfect match and mismatch hybridization intensities.

1   4.   The method of claim 1 wherein said relative hybridization
2   intensities are evaluated as differences of perfect match and mismatch hybridization
3   intensities.

1   5.   In a computer system, a method for analyzing a sample nucleic
2   acid sequence, said method comprising:
3   inputting a plurality of hybridization intensities of probes exposed to said
4   sample nucleic acid sequence, said plurality of hybridization intensities forming a
5   hybridization spectrum of said sample nucleic acid sequence; and
6   comparing said hybridization spectrum of said sample nucleic acid
7   sequence to a reference hybridization spectrum to obtain an indication of similarity.

1   6.   The method of claim 5 wherein said reference hybridization

2 spectrum comprises a plurality of hybridization intensities of probes from a chip exposed

3 to a reference nucleic acid sequence and said indication of similarity indicates similarity

4 between said sample nucleic acid sequence and said reference nucleic acid sequence.

1   7.   The method of claim 5 wherein comparing comprises:

2   applying a linear regression procedure to said hybridization spectrum of

3 said sample nucleic acid sequence and said reference hybridization spectrum.

1   8.   The method of claim 7 wherein said indication of similarity

2 .comprises a regression coefficient resulting from said linear regression procedure.

1   9.   The method of claim 5 further comprising:

2   selecting probes for said chip to minimize a number of probes needed to

3 ascertain similarity between said hybridization spectrum of said sample nucleic acid

4 sequence and said reference hybridization spectrum.

1   10.   The method of claim 5 wherein said probes are on one chip.

1   11.   The method of claim 5 wherein said probes are duplicated on

2 multiple chips and inputting comprises averaging hybridization intensities over multiple

3 chips to obtain said hybridization spectrum.

1   12.   The method of claim 9 further comprising selecting probes for said

2 chip to use a maximum number of probes wherein adding further probes does not

3 substantially improve performance.

1   13.   The method of claim 5 wherein said indication of similarity

2 indicates expression of a particular gene or expressed sequence tag (EST).

1   14.   The method of claim 5 wherein said indication of similarity detects

2 a mutation.

1   15.   In a computer system, a method for grouping genes into families,

2   said method comprising:

3   inputting a plurality of hybridization intensities of probes exposed to a

4   sample nucleic acid sequence, said plurality of hybridization intensities forming a

5   hybridization spectrum of said sample nucleic acid sequence;

6   repeating said inputting step for a plurality of sample acid sequences to

7   obtain a plurality of hybridization spectra; and

8   identifying similar ones of said hybridization spectra as indicating

9   expression of genes and ESTs that form a part of families.

1   16.   In a computer system, a method for analyzing a nucleic acid

2   sequence comprising the steps of:

3   inputting a plurality of hybridization intensities of probes exposed to said

4   sample nucleic acid sequence; and

5   applying a non-linear filter to said plurality of hybridization intensities.

1   17.   The method of claim 16 wherein said probes comprise pairs of

2   perfect match and mismatch probes, each pair including a perfect match probe perfectly

3   complementary to a subsequence of said nucleic acid sequence and a mismatch probe

4   having at least one base mismatch with said particular subsequence.

1   18.   The method of claim 17 wherein applying a non-linear filter

2   comprises:

3   for a particular base position, averaging hybridization intensities of perfect

4   match probes over subseqences aligned to said particular base position and subsequences

5   aligned to surrounding base positions, excluding outlying hybridization intensities.

1   19.   The method of claim 18 wherein applying a non-linear filter

2   comprises:

3   for a particular base position, averaging hybridization intensities of

4   mismatch probes over subseqences aligned to said particular base position and

5   subsequences aligned to surrounding base positions, excluding outlying hybridization

6   intensities.

1   20.   The method of claim 17 further comprising:

2   evaluating relative hybridization intensities for a plurality of subsequences

3   by comparison of hybridization intensities between perfect match and mismatch probes in

4   individual ones of said pairs; and

5   wherein applying a non-linear filter comprises for a particular base

6   position, averaging relative hybridization intensities over subsequences aligned to said

7   particular base position and subsequences aligned to surrounding base positions,

8   excluding subsequences having outlying relative hybridization intensities.

1   21.   The method of claim 17 wherein applying a non-linear filter

2   comprises:

3   for a particular base position, obtaining a median of hybridization

4   intensities of perfect match probes of subseqences aligned to said particular base position

5   and subsequences aligned to surrounding base positions, excluding outlying hybridization

6   intensities.

1   22.   The method of claim 18 wherein applying a non-linear filter

2   comprises:

3   for a particular base position, obtaining a median of hybridization

4   intensities of mismatch probes of subseqences aligned to said particular base position and

5   subsequences aligned to surrounding base positions, excluding outlying hybridization

6   intensities.

1   23.   The method of claim 17 further comprising:

2   evaluating relative hybridization intensities for a plurality of subsequences

3   by comparison of hybridization intensities between perfect match and mismatch probes in

4   individual ones of said pairs; and

5   wherein applying a non-linear filter comprises for a particular base

6   position, obtaining a median of relative hybridization intensities of subsequences aligned

7   to said particular base position and subsequences aligned to surrounding base positions,

8   excluding subsequences having outlying relative hybridization intensities.

1   24.   A computer-implemented method for detecting expression of a

2   plurality genes or expressed sequence tags (ESTs) comprising:

3   selecting probes to detect said plurality of genes or ESTs wherein each

4   gene or EST has an associated set of probes having a unique hybridization pattern when

5   exposed to sample having its gene or EST expressed; and

6   constructing an array of said selected probes.

1   25.   The method of claim 24 further comprising the step of:

2   exposing said selected probes to a target sample;

3   inputting a plurality of hybridization intensities of a set of probes

4   associated with a selected gene or ESTs, said plurality of hybridization intensities

5   forming a hybridization spectrum of said target sample; and

6   comparing said hybridization spectrum of said target sample to a reference

7   hybridization spectrum to obtain an indication of expression of said selected gene or EST.

1   26.   A computer program product for analyzing a sample nucleic acid

2   sequence, said product comprising:

3   code for inputting a plurality of hybridization intensities of pairs of perfect

4   match and mismatch probes, each pair including a perfect match probe perfectly

5   complementary to a particular nucleic acid subsequence indicative of expression of a gene

6   or EST and a mismatch probe having at least one base mismatch with said particular

7   subsequence;

8   code for evaluating relative hybridization intensities for a plurality of

9   subsequences by comparison of hybridization intensities between perfect match probes

10   perfectly complementary to said subsequences and mismatch probes having at least one

11   base mismatch with said particular subsequences;

12   code for applying a non-linear filter to said relative hybridization

13   intensities; and

14   a computer-readable medium for storing the codes.

1   27.   The product of claim 26 wherein the code for applying comprises:

2   code for, for a particular base position, averaging relative hybridization

3   intensities over subsequences aligned to said particular base position and subsequences

4   aligned to surrounding base positions, excluding subsequences having outlying relative

5   hybridization intensities.

1   28.   The product of claim 26 wherein said relative hybridization

2   intensities are evaluated as ratios of perfect match and mismatch hybridization intensities.

1   29.   The product of claim 26 wherein said relative hybridization

2   intensities are evaluated as differences of perfect match and mismatch hybridization

3   intensities.

1   30.   In a computer system, a product for analyzing a sample nucleic

2   acid sequence, said product comprising:

3   code for inputting a plurality of hybridization intensities of probes exposed

4   to said sample nucleic acid sequence, said plurality of hybridization intensities forming a

5   hybridization spectrum of said sample nucleic acid sequence; and

6   code for comparing said hybridization spectrum of said sample nucleic

7   acid sequence to a reference hybridization spectrum to obtain an indication of similarity;

8   and

9   a computer-readable medium for storing the codes.

1   31.   The product of claim 30 wherein said reference hybridization

2   spectrum comprises a plurality of hybridization intensities of probes from a chip exposed

3   to a reference nucleic acid sequence and said indication of similarity indicates similarity

4   between said sample nucleic acid sequence and said reference nucleic acid sequence.

1   32.   The product of claim 30 wherein comparing code comprises:

2   code for applying a linear regression procedure to said hybridization

3   spectrum of said sample nucleic acid sequence and said reference hybridization spectrum.

1   33.   The product of claim 32 wherein said indication of similarity

2   comprises a regression coefficient resulting from said linear regression procedure.

1   34.   The product of claim 30 further comprising:

2   code for selecting probes for said chip to minimize a number of probes

3   needed to ascertain similarity between said hybridization spectrum of said sample nucleic

4   acid sequence and said reference hybridization spectrum.

1   35.   The product of claim 30 wherein said probes are on one chip.

1   36.   The product of claim 30 wherein said probes are duplicated on

2   multiple chips and said inputting step comprises averaging hybridization intensities over

3   multiple chips to obtain said hybridization spectrum.

1   37.   The product of claim 34 further comprising selecting probes for

2   said chip to use a maximum number of probes wherein adding further probes does not

3   substantially improve performance.

1   38.   The product of claim 30 wherein said indication of similarity

2   indicates expression of a particular gene or expressed sequence tag (EST).

1   39.   The product of claim 30 wherein said indication of similarity

2   detects a mutation.

1   40.   In a computer system, a product for grouping genes into families,

2   said product comprising:

3   code for inputting a plurality of hybridization intensities of probes exposed

4   to a sample nucleic acid sequence, said plurality of hybridization intensities forming a

5   hybridization spectrum of said sample nucleic acid sequence;

6   code for repeatedly applying said inputting code for a plurality of sample

7   acid sequences to obtain a plurality of hybridization spectra;

8   code for identifying similar ones of said hybridization spectra as

9   indicating expression of genes and ESTs that form a part of informative families; and

10   a computer-readable medium for storing the codes.

1   41.   In a computer system, a product for analyzing a nucleic acid

2   sequence comprising the steps of:

3   code for inputting a plurality of hybridization intensities of probes exposed

4   to said sample nucleic acid sequence;

5   code for applying a non-linear filter to said plurality of hybridization

6   intensities; and

7   a computer-readable medium for storing the codes.

1   42.   The product of claim 41 wherein said probes comprise pairs of

2   perfect match and mismatch probes, each pair including a perfect match probe perfectly

3   complementary to a subsequence of said nucleic acid sequence and a mismatch probe

4   having at least one base mismatch with said particular subsequence.

1   43.   The product of claim 42 wherein said applying a non-linear filter

2   code comprises:

3   code for, for a particular base position, averaging hybridization intensities

4   of perfect match probes over subseqences aligned to said particular base position and

5   subsequences aligned to surrounding base positions, excluding outlying hybridization

6   intensities.

1   44.   The product of claim 43 wherein said applying a non-linear filter

2   code comprises:

3   code for, for a particular base position, averaging hybridization intensities

4   of mismatch probes over subseqences aligned to said particular base position and

5   subsequences aligned to surrounding base positions, excluding outlying hybridization

6   intensities.

1   45.   The product of claim 42 further comprising:

2   code for, evaluating relative hybridization intensities for a plurality of

3   subsequences by comparison of hybridization intensities between perfect match and

4   mismatch probes in individual ones of said pairs; and

5   wherein said applying a non-linear filter code comprises code for, for a

6   particular base position, averaging relative hybridization intensities over subsequences

7   aligned to said particular base position and subsequences aligned to surrounding base

8   positions, excluding subsequences having outlying relative hybridization intensities.

1   46.   The product of claim 42 wherein said applying a non-linear filter

2 code comprises:

3   code for, for a particular base position, obtaining a median of

4 hybridization intensities of perfect match probes of subseqences aligned to said particular

5 base position and subsequences aligned to surrounding base positions, excluding outlying

6 hybridization intensities.

1   47.   The product of claim 43 wherein said applying a non-linear filter

2 code comprises:

3   code for, for a particular base position, obtaining a median of

4 hybridization intensities of mismatch probes of subseqences aligned to said particular

5 base position and subsequences aligned to surrounding base positions, excluding outlying

6 hybridization intensities.

1   48.   The product of claim 42 further comprising:

2   code for, evaluating relative hybridization intensities for a plurality of

3 subsequences by comparison of hybridization intensities between perfect match and

4 mismatch probes in individual ones of said pairs; and

5   wherein said applying a non-linear filter code comprises code for, for a

6 particular base position, obtaining a median of relative hybridization intensities of

7 subsequences aligned to said particular base position and subsequences aligned to

8 surrounding base positions, excluding subsequences having outlying relative
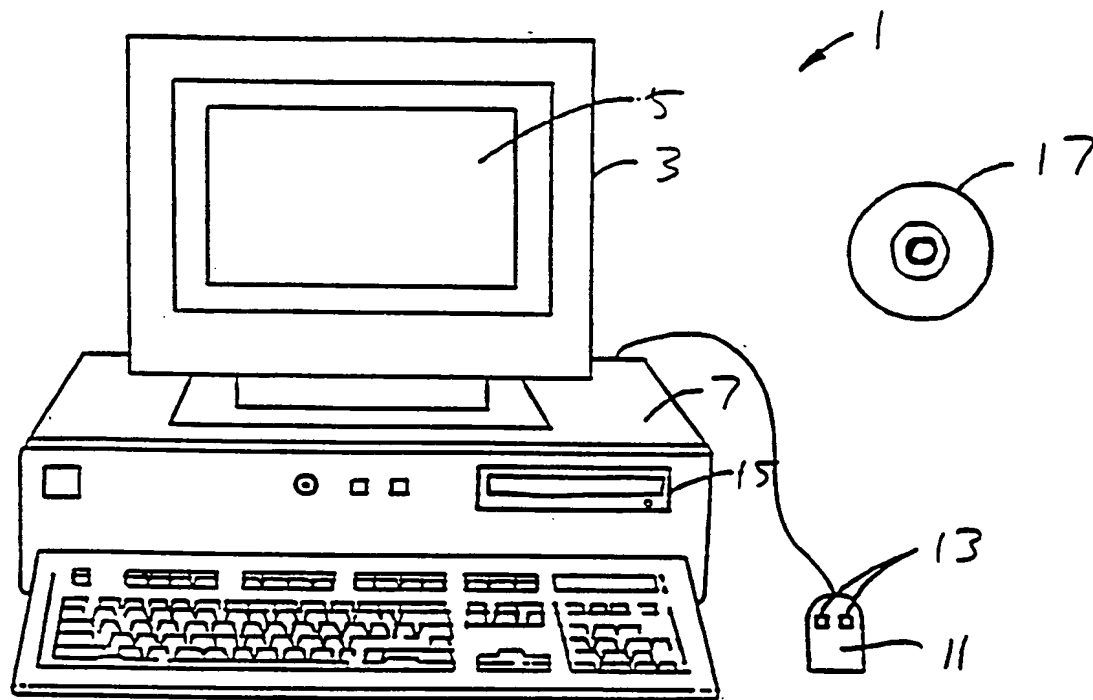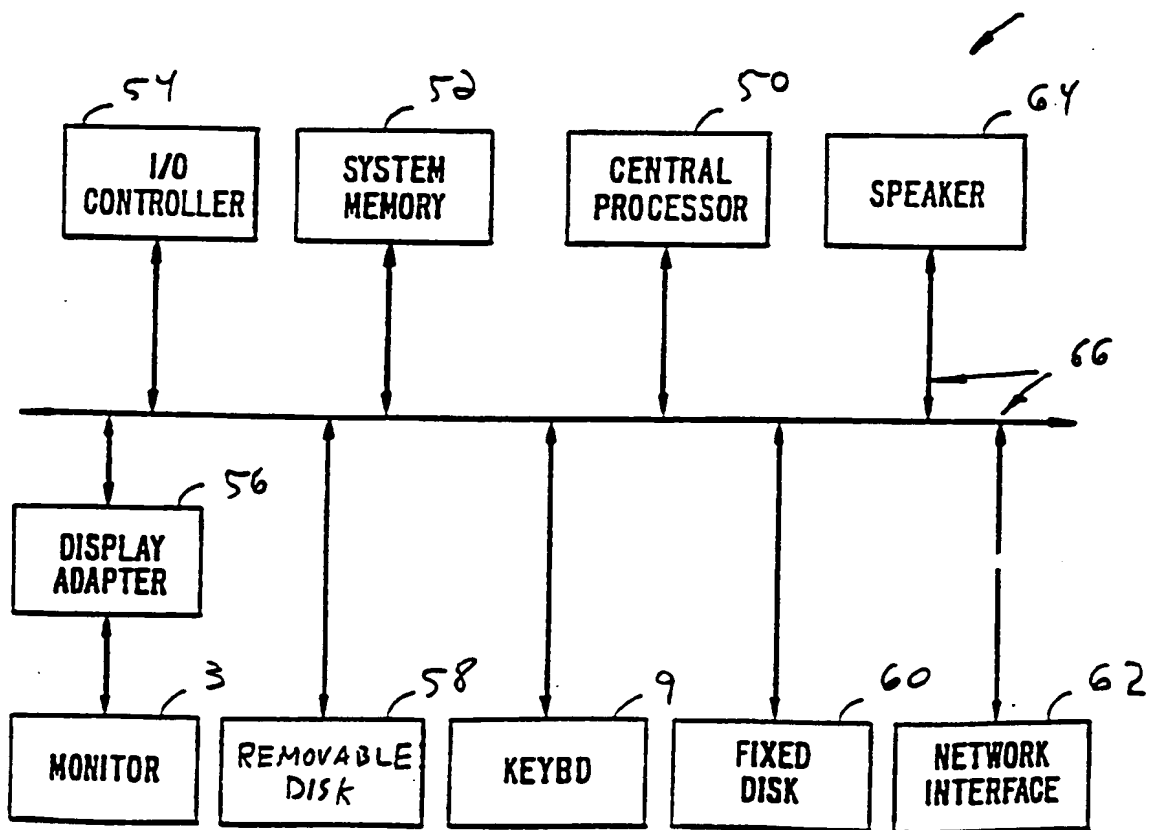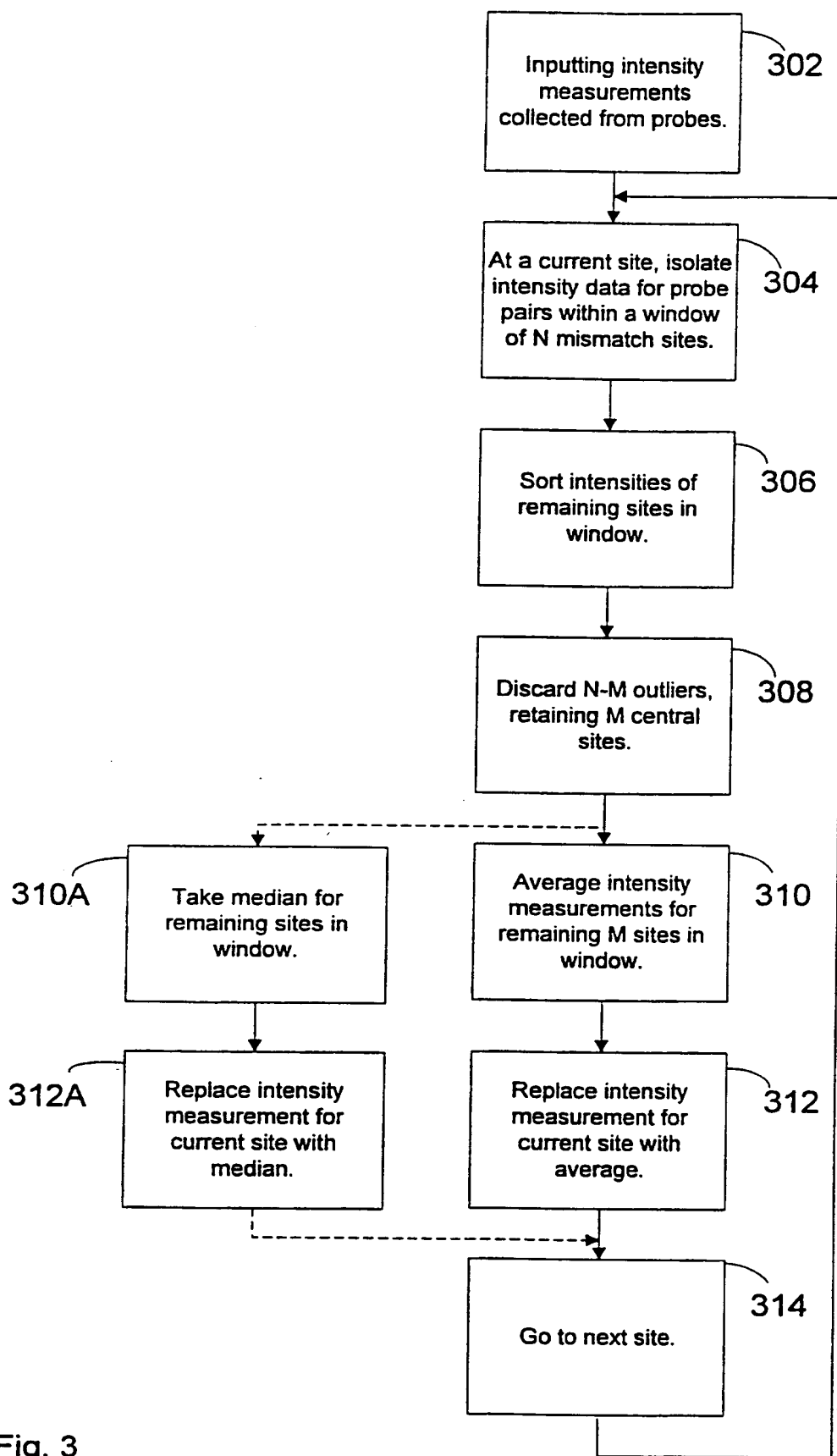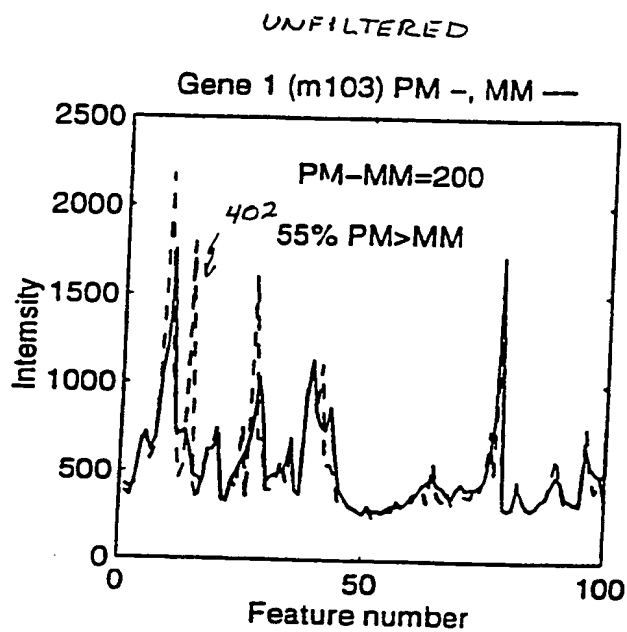
9 hybridization intensities.

FIG. 1

FIG 2

Fig. 3
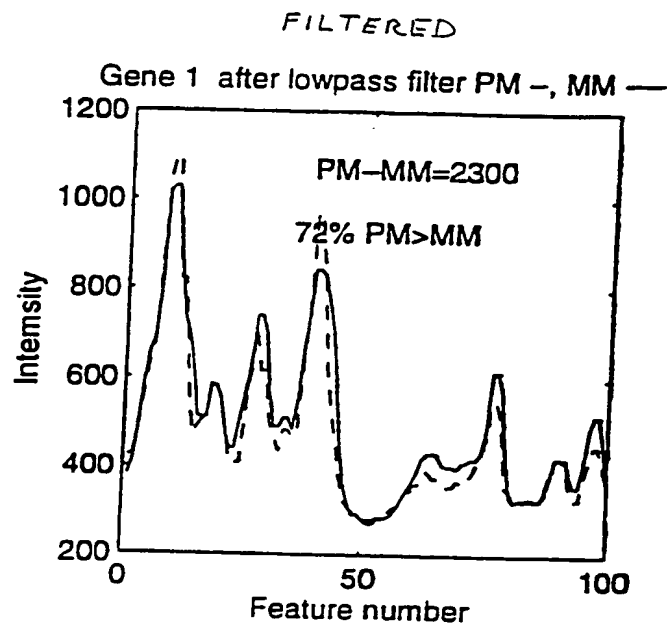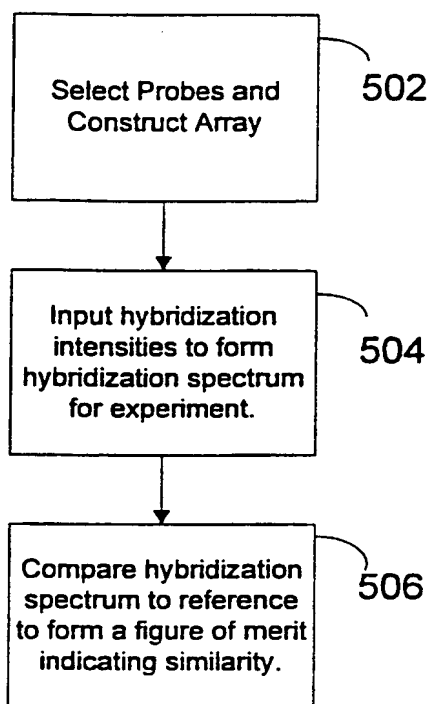
UNFILTERED

FIG 4A

FILTERED

FIG 4B

Fig. 5

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :C12Q 1/68
US CL :435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 422/50,68.1,69,82.05; 435/6,283.1,286.1,287.1,287.2; 708/801,805

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS, EMBASE, BIOTECH ABS, MEDLINE, AND WPI covering terms: hybridization, intensity, compare, code, non-linear, filter

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	WO 89/10977 A1 (ISIS INNOVATION LIMITED) 16 November 1989, see the abstract and sections 4-5 on pages 5-12.	24, 25 --------- 1-48
X --- Y	US 5,653,939 A (HOLLIS et al.) 05 August 1997, see the abstract and Figures 17 and 21 and associated text.	24, 25 --------- 1-48
X --- Y	US 5,800,992 A (FODOR et al.) 01 September 1998, see the abstract and column 51, line 31, through column 52, line 16.	24, 25 --------- 1-48

X	Further documents are listed in the continuation of Box C.		☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
02 FEBRUARY 2000	16 FEB 2000

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	ARDIN MARSCHEL
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)★

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 5,700,637 A (SOUTHERN) 23 December 1997, see the entire disclosure.	24, 25 ---------- 1-48
X --- Y	US 5,695,934 A (BRENNER) 09 December 1997, see the abstract, Figures 3 and 5, and column 16, line 24, through column 19, line 19.	24, 25 ---------- 1-48
X --- Y	US 5,795,716 A (CHEE et al.) 18 August 1998, see the abstract, Figure 8, and column 17, line 7, through column 19, line 28.	1, 5, 6, 10, 11, 14, & 16-25 ---------------- 1-48